Using AI Agents to Generate Recommendations Based on ACR Whitepapers

Paulo Kuriki, MD, Assistant Professor in Neuroradiology, Radiology, UT Southwestern Medical Center
Yin Xi, PhD; Ali Tejani, MD; Fernando Kay, MD, PhD; Krishna Sai Balineni, MS; Yee Ng, MD

Background/Problem Being Solved

The American College of Radiology (ACR) whitepapers provide evidence-based guidelines to aid radiologists in recommending actions for incidental findings, such as adrenal, liver, and pulmonary nodules. However, consulting these documents during routine practice is time-consuming and labor-intensive. While models like GPT-4 may theoretically incorporate this knowledge, their use introduces concerns related to privacy, hallucinations, and errors in extracting accurate information. This study investigates the application of AI agents to automate the extraction of adrenal nodule features and generate recommendations based on ACR criteria.

Intervention(s)

An agent-driven recommendation system was developed using Ollama and Llama 3.3 70B. A total of 769 Abdomen MRI reports were extracted. Using a separate dataset, prompt engineering was performed to optimize the extraction of structured JSON responses containing features to generate ACR-based recommendations. A Python function was developed to receive a JSON input and return an ACR-based recommendation. This function was integrated into Ollama as a tool, enabling it to generate recommendations upon request. The LLM was instructed to identify and extract adrenal nodule features and invoke the tool function to create structured recommendations. Results were manually reviewed by a 6-year experienced board-certified radiologist.

Barriers/Challenges

This pilot study was limited to reports from only two institutions. The use of a 70B model may pose computational challenges, while smaller models might achieve comparable results. Advanced prompt engineering techniques could further improve performance. Future efforts will focus on testing and fine-tuning smaller models to reduce computational demands.

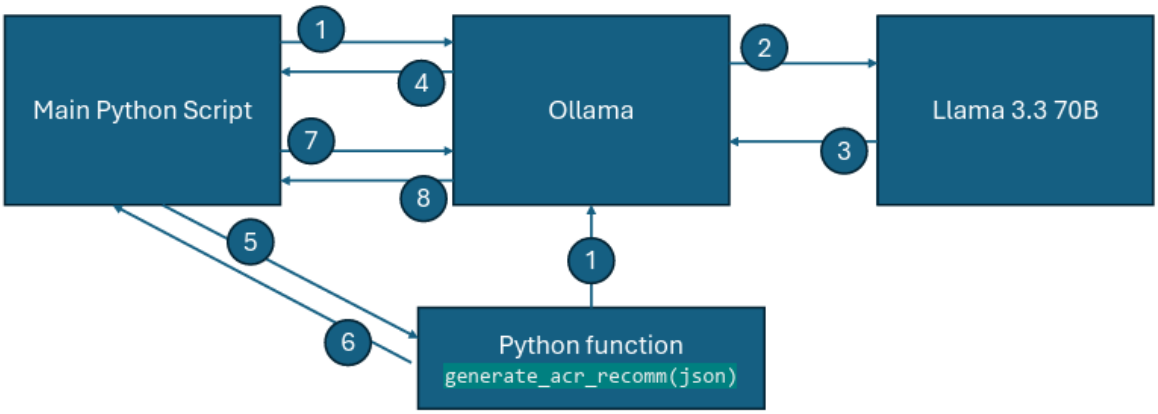
Outcome

From 769 reports, 98 recommendations for adrenal nodules were generated. Of these, 79 (80%) were classified as correct. Errors in the remaining 19 cases were attributed to failures in correctly extracting features, especially fat content, stability, or cancer history, when not explicitly described. Additional failures arose from issues in the tool calling process or incorrect function classification. Users anticipate substantial time savings by avoiding consultation of whitepapers, with radiologists unfamiliar with abdominal imaging benefiting most from ACR-aligned recommendations. These results demonstrate the potential for integrating Agentic AI into existing report generation workflows.

Conclusion/Statement of Impact/Lessons Learned

AI agents can significantly improve the generation of systematic, consistent, and evidence-based recommendations in radiology reporting. By leveraging external APIs, databases, and decision-making tools, these agents augment the capabilities of LLMs, improving accuracy and reproducibility. This study highlights the significant potential of agentic AI models to support more efficient and precise report generation, offering promising opportunities for integration into clinical workflows.

Figure(s)



1. A Python script sends a request to Ollama to identify and generate recommendations for any adrenal nodule in a report. In the same request, a reference to a Python function for generating ACR-based recommendations is provided as a tool.
2. Ollama asks Llama 3.3 70B to generate a response, along with instructions to call tools, if necessary.
3. The model generates as a response the extracted nodule features (JSON) and the tool to be invoked.
4. The response, containing the features in JSON and the function to be called, is returned.
5. The script calls the function **locally** with the JSON provided by the LLM as a parameter.
6. The function returns an ACR recommendation.
7. The script calls Ollama again, providing the initial request, extracted features, the requested function call, and the function result.
8. Finally, Ollama requests the model to craft a recommendation based on the ACR criteria.

Figure 1. This diagram details the data flow between the user, the AI Agent (in this case, a classifier function), Ollama, and the Llama model. AI Agents can potentially perform more complex tasks, such as retrieving data from EHR, downloading images, sending messages, generating billing, etc., allowing the model to expand its ability to take actions and solve complex problems.

Keywords

Artificial Intelligence/Machine Learning; Clinical Workflow & Productivity; Emerging Technologies; Patient/Family Experience